Learning and Inference in Computational Systems Biology

Wit, Ernst

Published in:
Biometrics

Publication date:
2012

Biometrics 68, 327–330
March 2012

BOOK REVIEWS

Introduction to General and Generalized Linear Models
(H. Madsen and P. Tyregod) *Clarice Demétrio*

Sample Sizes for Clinical Trials
(S. A. Julious) *Janet Wittes*

Learning and Inference in Computational Systems Biology.
(N. D. Lawrence, M. Girolami, M. Rattray and G. Sanguinetti) *Ernst Wit*

Brief Reports by the Editor

Applied Probability, 2nd edition.
(K. Lange)

Regression Estimators: A Comparative Study, 2nd edition.
(M. H. J. Gruber)

Common Statistical Methods for Clinical Research with SAS Examples, 3rd edition.
(G. A. Walker and J. Shostak)

MADSEN, H. and TYREGOD, P. **Introduction to General and Generalized Linear Models.** CRC Press, Boca Raton, Florida, 2011. xiii + 302 pp. $83.95/39.99. ISBN 9781420091557.

As stated by the authors, "this book contains an introduction to general and generalized linear models using ... likelihood techniques. The aim is to provide a flexible framework for the analysis and model building using data of almost any type."

The book is organized in seven chapters and two appendices. It begins with a general chapter presenting some motivating examples and a first view on the type of models that will be treated in the subsequent chapters. Chapter 2 presents a review of some basic concepts and definitions linked to the likelihood theory with examples, establishing the notation to be used, which I find is very useful. The presentation of general (Chapter 3) and generalized linear models (Chapter 4) is provided applying the likelihood methods, allowing for a clear comparison between the models. The concept of Gaussian mixed models is presented and developed in Chapter 5 and extended to hierarchical generalized linear models in Chapter 6. Exercises are proposed at the end of each chapter and some problems inspired by real life situations are considered in Chapter 7. The Appendices A and B summarize the delta method used to derive approximate expressions for the moments of a random variable and some probability distributions. All examples can be verified by the reader through the packages and code in R, and the analysis can be explored in great detail.

A minor point is that the authors use nonstandard terminology in a few places. For example, the chapter called "The likelihood principle" should probably have been called "The likelihood method", because the term "likelihood principle" normally refers to inference methods that depend on the likelihood function only, and not, for example, on the assumed distribution of the data, which is not what the authors have in mind.

This book is targeted to undergraduates in statistics but can be used by researchers as a reference manual as well. It is well written, easy to read and the discussion of the examples is clear. As a complement there is a collection of slides for an introductory course on general, generalized, and mixed effects models in the homepage cited in the preface of this book. This book has a good set of references but I missed citations to the paper by Jorgensen (1987) and to the book by Jorgensen (1997) when using the terminology "exponential dispersion family" created by Jorgensen (1987). In summary, I recommend this book as one of the textbooks to be discussed in a course for model building.

References

Jørgensen, B. (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society, Series B* **49,** 127–162.

Jørgensen, B. (1997). *The Theory of Dispersion Models.* London: Chapman & Hall.

Clarice G. B. Demétrio
Departamento de Ciências Exatas
Escola Superior de Agricultura "Luiz de Queiroz"
Universidade de São Paulo, São Paulo, Brazil

JULIOUS, S. A. **Sample Sizes for Clinical Trials.** Chapman & Hall/CRC, Boca Raton, Florida, 2010. Xxvii + 299 pp. $87.95/55.99, ISBN 9781584887393.

Clinical researchers who are inexperienced in dealing with statisticians often think of us as machines for calculating sample sizes: they come to us with a synopsis of a clinical trial and ask us to fill in the sections on sample size and statistical methods. Or, worse, they come to us with a sample size and ask us to justify it. Thus, a thoughtful book about sample size for clinical trials, directed at researchers, can serve as a

didactic tool as well as handbook for appropriate formulas. Steven A. Julious, a member of the faculty of the Medical Statistics Group at the University of Sheffield in England, has written extensively on design and analysis of clinical trials. I have always found his papers to be clear and interesting. His book, *Sample Sizes for Clinical Trials*, adopts his typical engaging style: he clearly has strong opinions and actually cares for his audience. He disarmingly states in his preface, "...this book is a little intentionally dry to enable the quick finding of an appropriate formula, application of a formula and a worked example." No, the book is not dry: it is lively, peppered with opinions and advice, and full of methods.

The fifteen chapters in the book are tightly organized. The first chapter, an introduction, briefly presents necessary background on the normal distribution and the central limit theorem. It then describes several types of trials, which will form the building blocks for the remainder of the book: superiority trials, equivalence trials, noninferiority trials, as-good-or-better trials (which I could not differentiate from noninferiority trials), bioequivalence trials, and trials designed to provide estimates with a given precision. Chapter 2, "Seven Key Steps to Cook up a Sample Size," provides a template for the remaining chapters; it describes the ingredients one needs to calculate a sample size and gives some worked examples. The remaining chapters are organized by distribution of the data (normal, binary, ordinal, and time-to-event) and type of trial, including crossover designs as well as the types listed above. Each chapter gives a series of formulas for calculating sample size under different conditions, some worked examples, some discussion, and, when appropriate, some warnings. Several of the warnings show pitfalls into which researchers often fall. The section on how to calculate variance of change from published data was particularly incisive. Julious points out that many people use the variances from baseline tables of papers rather than the data from the analysis of variance, which accounts for the correlation between baseline and follow-up. Use of the variance at baseline implicitly assumes a correlation of 0.5; many variables of interest in clinical trials have correlations of 0.6 to 0.8.

But there were peculiar lapses. Julious calls the Central Limit Theorem the "law of large numbers," but the two theorems are not the same; the former describes the conditions under which a mean has a normal distribution whereas the latter speaks to the convergence of a sample mean to the true mean. For the intended audience of this book, this error is not important because the clinical researcher is likely to gloss over technical statements of probability theorems. But for the statistician, it is a troubling statement. More important lapses are those related to practical issues. As I was reading the book, I heard Dr. Seuss chanting in the background, "It's a pretty good zoo...but if I ran the zoo, I'd make a few changes...that's what I'd do." So, here are three changes I hope Julious will address in his next edition.

First, the book should have dealt much more seriously with the problems and implications of dropouts. The only reference to dropout is in Chapter 3, which simply tells the reader to adjust sample size by dividing the calculated sample size by $(1 - p)$, where p is the proportion dropping out. This approach is inadequate, for it implicitly assumes that the analysis of the study included only completers, which may give quite biased results. Given the large effect dropout can have on results of clinical trials if the trials are analyzed appropriately (see *The Prevention and Treatment of Missing Data in Clinical Trials,* [Panel on Handling Missing Data in Clinical Trials, National Research Council, 2010]), recommending this naive method of calculating sample size in the presence of dropout can lead to a sample size that is seriously too low for the desired power.

Second, the book would be much more useful if it were not limited to designs with two treatment groups. Phase 2 trials often have several dose groups and frequently Phase 3 trials study two doses and a control. Many noninferiority trials have three groups: test product, active control, and placebo control. Dealing with more than two groups adds complexity to the design of a study and to the calculation of sample size because of the necessity of coming to grips with multiplicity. I fear that the typical clinical researcher may not understand how to move from two to more than two treatment arms and may simply think that for equal allocation one has to multiply the sample size by $g/2$ where g is the number of groups.

Third, the book would be more useful if it contained a fuller description of methods for sample size calculation in survival data. Reference to the work of Schoenfeld (1983) and Lakatos (1988), for example, would have been valuable.

In summary, the book has many useful formulas for a variety of designs. It is well organized, so users can easily find the section relevant to their needs. And the plethora of worked examples is very helpful. But I would not recommend this book for someone designing a survival study or for someone designing a study where dropout is likely to occur. I see it as a useful introduction to the clinical researcher and as a reference for the statistician interested in sample size formulae for specific designs.

References

Lakatos, E. (1988). Sample size based on the log-rank statistics in complex clinical trials. *Biometrics* **44,** 229–241.

Schoenfeld, D. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* **39,** 499–503.

Panel on Handling Missing Data in Clinical Trials, National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press.

Janet T. Wittes
Statistics Collaborative, Inc.
Washington, DC

LAWRENCE, N. D., GIROLAMI, M., RATTRAY, M., and SANGUINETTI, G. **Learning and Inference in Computational Systems Biology.** The MIT Press, Cambridge, MA, 2010. Xxv + 375 pp. $42.00/28.95. ISBN 9780262013864.

Systems biology is one of the most challenging inferential fields currently around. Despite the highly structured underlying biological system, the data are typically noisy, the system is high-dimensional, and replicates are few. These challenges

lead to a variety of functional data analysis questions, which this edited volume tries to address.

I have to start by saying that I generally dislike edited volumes. Too often do they disguise a wild collection of articles under a supposedly unifying title. Fortunately, this collection is different. The editors put the book together around a central theme and purpose, to wit statistical inference of genomic systems, with mostly consistent notation throughout the book and with abundant cross-referencing between the various articles. The aim of the book is to elucidate biological processes from the point of view of the available data and in this way the edited volume has even marked advantages: the various authors bring in various methods to approach various types of data to answer, in essence, a single question. It would have, otherwise, been unlikely to find in the same book methods for dealing with nonlinear ODEs, SDEs, mixtures of factor analyzers, and a variety of other methods.

The editors of the book are four highly respected quantitative scientists at the very cutting edge of statistics, machine learning, and systems biology. They have taken their job as editors very seriously. They all wrote individual contributions as well as two introductory chapters to aid the flow of the book. Overall it is an attractive volume with many illustrative examples, clear typesetting and informative graphs.

The book has a predominantly Bayesian flavor. Lawrence argues in the Introduction that this is an ideal framework "to sustain multiple hypotheses" in Dawid's prequential sense of competing predictive hypotheses. However, most of the contributions in the book can be labeled under Senn's you-may-believe-you-are-a-Bayesian-but-you-are-probably-wrong (Senn, 2011). Rarely do any of the authors pay serious attention to the priors on the models (i.e., hypotheses) or even on the parameters. Bayesianism in this book is merely an operational choice and justification does not go much beyond statements such as "Bayesian estimators [sic] are self-justified (they minimize Bayesian risk)" (p. 70). In my mind, the frequentist framework, where hypotheses can simply live side by side without having to be integrated, seems more consistent with Popper's and Dawid's ideas. However, I do not have major objections to the exclusive focus on Bayesian machinery, also because the contributions are from some of the finest authors in the field.

The mathematical rigor varies from author to author, though. The notation $X_t = f(X_t)$ at some place is used not as a fixed point, but as a functional form $X_i(t) = f(X_{-i}(t))$ and even as some ordinary differential equation $\mathrm{d}X_i/\mathrm{d}t = f(X_{-i})$ (p. 46). Elsewhere identifiability is defined as "given some data [sic] only one value of the parameter can produce the observed behaviour" (p. 69). All this is a bit sloppy. On the other hand, an attractive feature of the book is that authors do not shun sacrificing some rigor for readability. The book has a natural progress. More conceptual matters precede the technical ones and where necessary background information is given so that any graduate level statistician can follow the argument. The only exception is perhaps Chapter 12, which, like Chapters 10 and 11, also focuses on stochastic differential equation modeling for single cell data, but which is much more basic compared to either two.

The volume has its roots in a EU collaborative program to encourage the interaction between computational biologists, statisticians, modelers, and machine learners. The book itself does not quite fulfill that purpose. It is clear that this is not a handbook for a computational biologists. That it may serve as a foundation for a "decade of research in systems biology" (p. 7) may sound somewhat ambitious, but it is indeed an excellent book for a computational Ph.D. student (and their advisors) to pick up interesting ideas that can be developed further. As such it is an excellent volume for the general readership of Biometrics and the very mild price of the book will not and should not stop any interested reader from buying it.

Reference

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals.* **2,** 48–66.

Ernst Wit
Johann Bernoulli Institute
University of GroningenGroningen, The Netherlands

BRIEF REPORTS BY THE EDITOR

LANGE, K. **Applied Probability**, 2nd edition. Springer, New York, NY, 2010. Xvi + 436 pp. $89.95/€85.55, ISBN 9781441971647.

Applied Probability, first published in 2003, is a book that has already been established. A review report of the first edition of this book has been published in Biometrics by R. N. Bhattacharya (*Biometrics* 60: 562–572, 2004) and also by other reviewers in other journals. In its first edition, there were 13 chapters covering basics of probability; calculation of expectation; convexity, optimization, and inequalities; combinatorics and combinatorial optimization; Poisson processes; discrete-time and continuous-time Markov chains; branching processes, martingales; diffusion processes; Poisson approximation; and number theory. This new edition brings not only some additional material and subsections within its original chapters but also two new chapters on asymptotic and numerical methods, and an appendix with some more detailed mathematical theory. As in its first version, the new edition of *Applied Probability* stands out by its balance between theory and application, and its extensive list of exercises provided in each chapter. *Applied Probability* is a book that can be used as a textbook or a complement for an advanced graduate course in probability. In addition, as its chapters are stand-alone

discussions of topics that do not necessarily need to be followed in order, it is also an excellent reference book for researchers and advanced graduate students.

GRUBER, M. H. J. **Regression Estimators: A Comparative Study**, 2nd edition. The John Hopkins University Press, Baltimore, MD, 2010. Xii + 412 pp. $110.00, ISBN 9780801894268.

Regression Estimators: A Comparative Study was first published in 1990, bringing an ample discussion and comparison of different mathematical formulations of ridge-type regression estimators, including both frequentist and Bayesian methods. The second edition of the book is intended to include some new developments in the field of regression estimators, as well as to improve the overall volume on the basis of suggestions from reviewers. The book is divided into 5 parts, with a total of 14 chapters. Chapters 1 and 2 (Part I) discuss the need for alternatives to least squares estimators, provide a historical survey, and summarize basic ideas in matrix theory and statistical decision theory. Chapters 3 and 4 (Part II) present frequentist and Bayesian estimators, and investigate the mathematical relationships between them. Chapters 5 through 8 (Part III) discuss the efficiency of the estimators. Chapters 9 and 10 (first two chapters of Part IV) present applications of the methods to Kalman Filter and analysis of variance, respectively. These first 10 chapters are revised versions of the original chapters in the first edition of the book. The new material available in this second edition refers mostly to four new chapters added to it. Chapter 11 (third and last chapter of Part IV) discusses how penalized splines and ridge-regression estimators are related. Finally, Chapters 12 through 14 (Part V) present some new results and developments about the behavior of ridge-type estimators with respect to Zellner's balanced loss function, the efficiency of the estimators with respect to the LINEX loss function, and the measurement of the Rao distance between ridge estimators using some ideas of differential geometry. The book provides a respectable reference source of ridge-type regression estimators, which should be valuable for statisticians, researchers, and graduate students who use and study ridge regression and lineal models in general.

WALKER, G. A. and SHOSTAK, J. **Common Statistical Methods for Clinical Research with SAS Examples**, 3rd edition. SAS Press, Cary, NC, 2010. Xiv + 536 pp. $79.95, ISBN 9781607642282.

The book by Glenn Walker and Jack Shostak is evidently a cookbook, which has been deliberately written primarily for nonstatisticians involved in clinical research and interested in SAS applications as well. In addition, examples provided in the book may be useful also for applied statisticians new to clinical applications, or for new SAS users working in clinical research. This third edition of the book is an update of the content of its previous version, especially in regard to new SAS procedures such as PROC MIXED and PROC GLIMMIX, for linear and generalized linear mixed effects models, and ODS Graphics features available with SAS 9.2. The book starts with two chapters covering some basic statistical concepts and an overview of hypothesis testing. Other 21 chapters present various statistical hypothesis testing techniques, ranging from simple *t*-tests, to analysis of variance (ANOVA) techniques, including crossover designs and repeated measurements, to several nonparametric tests. Some more advanced modeling approaches are also discussed, such as logistic regression and Cox proportional hazards model. The book ends with a series of appendices with additional material on common distributions, ANOVA concepts, multiple comparison methods, among others. Overall, the book is well written and nicely organized, and successfully accomplishes its goal of providing a SAS-oriented guide for nonstatisticians regarding common statistical tests used in clinical research.